A Survey on Short Text Analysis in Web

Rafeeque P C
Dept. of Information Science and Technology
Anna University Chennai
Email: rafeeqpc@yahoo.co.in

Sendhilkumar S
Dept. of Information Science and Technology
Anna University Chennai
Email: thamaraikumar@cs.annauniv.edu

Abstract—With the recent explosive growth of Short text in the Internet and blog-sphere, Short text classification and analysis has been identified as a booming research topic in recent times. Short text classification is a challenge due to its sparse nature, noise words, syntactical structure and colloquial terminologies used. It is usually difficult for traditional similarity measures to detect intrinsic relationship among Short text snippets as they contain very limited common words. Although there are several reviews done on Text classification in general, there are no systematic reviews on Short text classification and analysis. This survey discusses the existing works on Short text analysis and the related issues and challenges. The effectiveness of these algorithms have been analysed by using standard analytical measures.

Index Terms—Short text, Classification, Clustering, Opinion mining

I. Introduction

The advent of web 2.0 and the rapid development of mobile communication has dramatically changed the way to express the feeling, attitude, mood, passion etc. in broad categories as fragment information unlike traditional lengthy articles. Short text exists in a variety of forms: SMS messages, review about various products, image captions, forum posts, Twitter messages(Tweets), Blog and news feeds, code snippets, Fre-quently Asked Questions(FAQ) etc. With the explosive growth of these Short texts it has become difficult for users to utilize information and to classify and catalog Short texts.

The information mentioned above is a rich and useful source for classifying news, classifying Short snippets returned by the Web search engines, identifying and removing erotic content messages, query classification, classifying/clustering tweets and blog messages, classifying similar questions/answers, clustering scientific abstracts, sentiment detection and emotion aware clustering etc. To achieve this it is necessary to analyse Short text snippets or messages.

Text classification(long text/document) using traditional methods has long being studied in the IR field as a means of improving retrieval efficiency. Most traditional techniques for measuring the similarity of two documents mainly focus on word frequency and provide enough word co-occurrences or shared context for good similarity measure. Because of the sparseness of Short text snippets, state-of-the-art techniques usually fail to achieve desired accuracy[9].

There are several additional issues that have to be dealt with Short text classification. First, feature selection and reduction is the most important step in text classification. But due to the data sparseness of Short text this has became a challenge. So additional methods need to be incorporated to inflate the Short text.

Second, semantic similarity between entities are important in many web related tasks like automatic annotation or classification of web articles, news, tweets etc. Moreover Semantic similarity between entities changes over time and across domains. For example, apple is frequently associated with computers on the Web. However, this sense of apple is not listed in most general-purpose thesauri or dictionaries [12]. Semantic similarity is hard to achieve due to the lack of content and context. Third, the description of Short text is con-cise and need not have regular grammar. Due to this, standard NLP techniques cannot achieve desired results. Forth, Short text classification should address the quantity of information, dynamic nature of the Internet and the availabilty of training data.

Most of the research works in recent times have primar-ily focused on addressing the above mentioned problems. This survey uses systematic literature review pertinent to the challenges associated with the Short text analysis. The effectiveness of the methods have been analysed based on the corpus used and by using analytical measures.

The rest of the paper is structured as follows. Section 2 discusses the general framework used for rviewing Short text analysis. section 3 discusses the different approaches used for Short text analysis. Section 4 examines the evaluation measures and common benchmarks used for Short text anal-ysis. Section 5 addresses the major issues associated with Short text analysis. Last section concludes with challenges and discussion on future work.

II. General Framework

The following framework was used as illustrated in Figure 1 for reviewing Short text analysis. Most of the prior work on Short text analysis has focused on data sparseness problem. One of the intuitive method that can be applied to eliminate this problem is to extend the sparse features of Short text with additional information to make it appear like a long text or document. After fetching the text from its source, which could be a database repository, blog, a file system or the World Wide Web, should be preprocessed for feature extraction. Then, Short text similarity measures, classification or clustering, sentiment detection, etc. can be applied.

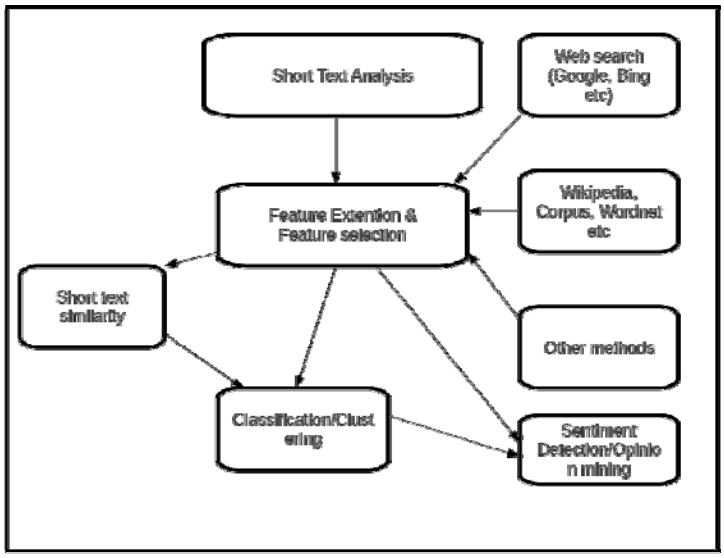

Fig. 1. General Framework for reviewing Short Text analysis

III. SHORT TEXT ANALYSIS - DIFFERENT APPROACHES

Different approaches used for Short text analysis were classified based on the following. (1) Additional sources or methods used for feature expansion. (2) Semantic similarity measures used. (3) Classification or clustering techniques used. The following section briefs few approaches used to find the similarity among Short text snippets.

A. Semantic similarity using web search for data enrichment

Semantic relationships between concepts or words can be used to discover inherent relationships between descriptions of entities. Semantic similarity is an important measure to classify or cluster Short text.

Shami and Heilman [8] introduced a new similarity measure called web kernel similarity function. Web-based similarity kernel is simply the inner-product of the expanded vector representations, denoted as QE(q) QE(s) where q and s are pair of Short text segments. Query expansion associated with each query is represented as a TF-IDF weighted term vector. Each vector is normalized and the centroid of the set of vectors is computed. Semantic similarity between two queries is then defined as the inner product between the corresponding centroid vectors. Yih and Meek [11] extended the Web-kernel similarity function by using relevance weighted inner-product of term occurrences rather than TF-IDF.

Bollegala et al. [12] integrated page count(the number of pages that contain the query words) based similarity score and Lexico-Syntactic patterns from snippets returned by search engines using Support Vector Machine to find the similarity score.

Quan et al. [9] used a novel method to find the Short text similarity. They first identified distinguished words from two Short text snippets and compared with the set of selected third-party topics to know the implicit relationship between the terms. They calculated the similarity by considering not only the common words but the relationship of the distinguishing terms also. Finally they applied the cosine similarity to the modified vectors.

Metzler et al. [10] extended the query by using web search results and they used 3 similarity measures(Lexical, Proba-bilistic and Hybrid) for query-query similarity matching.

Even though the above mentioned methods achieved good results in their particular domain, they are not suitable for real time applications as it is not feasible to perform semantic similarity search on every pair of Short text messages and it is time consuming. Moreover these techniques require additional word or entity disambiguation approaches. But these approaches can be used if there is no pre existing up-to-date taxonomies.

TABLE I
COMPARISON OF SEMANTIC SIMILARITY APPROACHES WHICH USED WEB SEARCH FOR THE FEATURE EXTENSION.

Author	Feature Extn	Domain	Dataset
Shami[8]	Query Expansion using WS	Query-Query similarity	Queries collected from Google
Yih[11]	Query related docs through WS	Query suggestions	Randomly selected query suggestions
Bollegala[12]	Snippets returned by WS .	Word-pair matching, Community mining	Miller charles Data set
Quan[9]	Probabilistic topics	Question categorization, Paraphrase matching	Questions, MS paraphrase
Metzler[10]	Expanded Queries Using WS.	Query-Query Similarity	Queries from MSN Search Query log

B. Semantic similarity using data repositories for data enrich-ment

Recent approaches on semantic similarity of Short texts utilized data repositories like Wikipedia as external data source to enhance the feature sparseness.

Okazaki et al. [6] used lexical database for data enrichment. They used lexical database to identify the lexical relationship between terms appearing in a sentence. Then sentence sim-ilarity is calculated by aggregating similarity values of all pairs of words. However, their method give more impact to a word having more synonyms and does not deal with how to choose the meaning of a word with polysemy from the lexical database.

Li et al. [1] also employed a lexical knowledge base and proposed a sentence similarity measurement based on lexical database and word ordering. For each sentence, a raw semantic vector is derived with the assistance of a lexical database. A word order vector is formed for each sentence, again using information from the lexical database. Finally, the sentence similarity is derived by combining semantic similarity and word order similarity.

Michale et al. [3] introduced another method for measuring the similarity of Short text snippets, which used both corpus-based and knowledge-based measures. They described two corpus-based and six knowledge-based measures of word semantic similarity, and showed how they can be used to derive a text-to-text similarity metric and they evaluated on a paraphrase recognition task.

Islam et al. [5] proposed a Semantic Text Similarity (STS) method which determines the similarity of two texts from semantic and syntactic information (in terms of common-word order) that they contain. They consider three similarity functions in order to derive a more generalized text similarityTABLE II

COMPARISON OF SEMANTIC SIMILARITY APPROACHES WHICH USED DATA REPOSITORIES(WORDNET, LEXICAL DATABASE ETC. FOR THE FEATURE EXTENSION.

Author	Feature Extn.	Domain	corpus
Okazaki[6]	Lexical dictionary	Text Summarization	Japanese Newspaper Article
Li[1]	Wordnet, Brown corpus	Sentence Similarity	30 selected sentence pairs
Mihalcea[3]	Wikipedia, Wordnet	Paraphrase recogn.	MS paraphrase corpus[45]
Islam[5]	Corpus	Sentence & paraphrase recogn.	30 sentence pair MS paraphrase corpus
Remage[4]	Wordnet, corpus statistics.	paraphrase recogn.	MS paraphrase corpus
Khaled[2]	Wordnet	Sentence similarity	30 sentence pair MS paraphrase corpus

method. First, string similarity and semantic word similarity are calculated and then used an optional common-word or-der similarity function to incorporate syntactic information. Finally, the text similarity is derived by combining string sim-ilarity, semantic similarity and common-word order similarity with normalization.

Remage et al. [4] used a random graph walk framework for measuring the semantic relatedness. Instead of comparing two bags-of-words directly, they compared the distribution of each text induces when used as the seed of a random walk over a graph derived from WordNet and corpus statistics.

More recent work by Khaled et al. [2] used word sense disambiguation and synonym expansion by using Wordnet to provide a richer semantic context to measure sentence similarity. For each of the sentences being compared, they first applied a word sense disambiguation step to identify the sense in which words are being used within the sentence. Semantic vetors are then estimated by applying synonym expansion on them. The similarity between semantic vectors can then be calculated using a standard vector space similarity measure such as cosine similarity.

C. Short text Classification

Formally text classification is a one-to-one or one-to-many mapping from the unclassified text to the given categories. Short text classification requires sufficient number of training examples to achieve the high accuracy. It is often both im-practical and extremely tedious and expensive to hand label a sufficient number of training examples to achieve the high accuracy that is needed for this task [13]. Many recent works on Short text classification addressed these issues. Following section briefs few approaches used to classify Short texts.

Zelikovitz et al. [13] used the unlabeled corpus as back-ground knowledge for the learner, to aid it in its decision task. Rather than directly comparing a new unlabeled example to elements of the labeled training corpus, they used theunlabeled background knowledge as a 'bridge', to connect the new example with labeled examples. Their approach is especially useful in cases with small amounts of training data and when each item in the data has few words. They have used LSI with back ground text [18] and Transductive Latent Semantic Indexing (LSI) [19] for text classification.

Pu et al. [14] proposed that Latent Semantic Analysis(LSA) is a good method for preprocessing and used text classifier based on Independent Component Analysis (ICA). Using ICA and LSA together rather than only using ICA in Chinese Short-text classification achieved better results.

Phan et al. [7] not only uses the explicit user defined categories in Wikipedia but extracts hidden topic of Wikipedia articles to gain more knowledge. They have introduced a new method to classify Short text snippets based on Latent Dirichlet Allocation[LDA] [44].

One of the important problem of enhancing the feature set by using external knowledge is the "Curse of Dimensionality". Sriram et al. [16] proposed a small set of domain-specific features extracted from the authors profile and text to classify tweets to a predefined set of generic classes such as News, Events, Opinions, Deals, and Private Messages.

Wei et al. [20] proposed association rule for Short text fea-ture extension and Faguo et al.[22] suggested rule and statistics to feature extraction. Liu et al. [21] used feature selection model based on parts of speech and HowNet(knowledge base for feature selection) for blog mining.

In contrast to a standard single-value classification where each instance is signed exactly one class label, multi-value (also called multi-label) classification allows for assigning an arbitrary number of labels to instances. Heßet al. [15] proposed a multi-value classification scheme.

D. Short text clustering

Unlike classification, clustering is a unsupervised method to group or partition objects without advance knowledge of the group definition. As the amount of Short text snippets generated in the web is very huge, users often face the problem of information overload. Clustering could be one of the solution to this problem. Short text clustering is considered to be complex due to the the low frequencies of vocabulary terms in Short texts. Custering will be harder if the domain is narrow (vocabulary overlapping level of the Short documents is very high)[31], [32]. Most of the prior works have focused these two domains. Following section briefs few approaches used to cluster Short texts.

Banerjee et al. [23] used titles of selected Wikipedia articles to augment the text quality and clustered Short text snippets based on the enriched representation. They used a snapshot of the Wikipedia database and all queries were directed to this database upon which the Lucene index was built.

Hu et al. [24] enhanced existing features by using both WordNet and Wikipedia. They exploited internal semantics to provide a deep understanding of the original Short texts and external semantics which incorporate the concepts derived from the world knowledge to reduce the semantic gap. They also proposed a novel hierarchical resolution phase to parse through the Short text and categorize them into segments, phrases and words. From this pool, seed phrases were carefully selected which formed the queries to Wikipedia and WordNet.

Zamir and Etzioni [25], [26] developed an interface to the results of the HuskySearch meta-search engine, which dynamically groups the search results into clusters labeled by phrases extracted from the snippets. They proposed a clustering method called Suffix Tree Clustering(STC). STC is a linear time clustering algorithm (linear in the size of the document set) that is based on identifying phrases that are common to groups of documents. In this method, common phrases shared by a set of documents are firstly identified by the Suffix Tree, and documents are then clustered according to these common phrases.

Hui He et al. [27] used N-gram feature extraction and RPCL (Rival Penalized Competitive Learning) [43] to cluster Chinese Short texts for mining web topics based on Chinese chunks. RPCL is a clustering algorithm used commonly for speech recognition and image segmentation clustering.

Yang Li and Rahi [33] have developed a tool, IntelliGrouper, to cluster search results returned from search engine. They clustered documents based on the phrases that contain the search words. Each document is searched for search words as well as phrases that contain the search words. Then the documents containing these phrases are clustered together.

Kummamuru et al. [28] propose a hierarchical clustering al-gorithm, DisCover, to organize search results into a hierarchy. They build a topic hierarchy for a collection of search results retrieved in response to a query.

Zeng et al. [34] introduced a new solution for clustering search results by reformalizing the clustering problem to a phrase ranking problem. Thus they convert an unsupervised clustering problem to a supervised learning problem. They manually rank the importance of all the n-grams in the search results corresponding to thirty queries. For a given query and the ranked list of search results, their method first parses the whole list of titles and snippets, extracts all possible phrases (n-grams) from the contents, and calculates several properties for each phrase such as phrase frequencies, document frequencies, phrase length, etc. A regression model learned from previous training data is then applied to combine these properties into a single score. However, labeling all the n-grams from the search result of a query is very time-consuming.

More recent work Ni et al. [36] proposed a new Short text clustering strategy, Termcut. They modeled collection of Short text snippets as a graph G = (V; E), where a vertex represents a text snippet and each weighted edge between two vertices measures the relationship between the two vertices. The core terms are found on the basis of minimizing a new criterion, RMcut. The RMcut criterion is a clustering quality criterion, which measures the quality of clusters according to the clustering principle “minimizing the inter-cluster similarity while maximizing the intra-cluster similarity”. Based on the TermCut strategy, they proposed two clustering algorithms CNTC(Cluster Number based TermCut) and TTC(Threshold based TermCut) to cluster Short text snippet for different applications. They have achieved good results on Question and Short snippet dataset.

Scientific abstracts is an another new domain in which clustering is being applied. Makagonov et al.[29] addressed the issue of clustering scientific abstracts and they used a new approach for selecting keywords from the word frequecy list. They considered objective criteria related to frequency of words with respect to general lexis and the expected number of clusters. They used a weighted combination of cosine and polynomial measure for finding the similarity between abstracts.

Clustering narrow domain Short texts(e.g., all abstracts only on Data mining or all on Computational liguistics) is hard due to the high vocabulary overlapping associated to narrow domains. Alexandrove et. al[30] proposed a new method by grouping keywords and using an adequate document similarity measure as mentioned in [29]. They used MajorClust method for clustering both keywords and documents.

Pinto et. al[31], [32] introduced novel measures to automat-ically determine whether corpus is made up of narrow domain short texts or not. They proposed a domain-independent self-term expansion methodology to enrich baseline corpus by adding co-related terms from an automatically constructed lexical knowledge resource obtained from the same target data set (and not from an external resource). They used this technique to cluster scientific abstrats which belongs to narrow domain.

E. Opinion mining from Short text

Today millions of comments and responses are expressed in blogshere, forums, social media and social networking sites etc. in the form of product reviews, political viewpoints, tweets etc. Due to the micro blogging people often express their views as Short messages. Opinion mining has been in the research for the past 10years. The issues and challenges of opinion mining has reviewed extensively in many papers[37], [38], [39], [40], [41]. Opinions expressed in the Short text form pose new challenges as they often irregular in grammar and cryptic in nature.

Opinion mining from Short texts has many potential appli-cations

such as attitude analysis from tweets, on line message filtering, SMS sentiment classification, opinion summarization from Short product reviews etc. Recent work by Thelwall M et. al[42] used new algorithm, SentiStrength, to extract sentiment strength from informal English text. They evaluated the algorithm by using MySpace comments and achieved fairly good result.

As Internet is overwhelmed with Short text opinions, in future more work is needed to explore data mining and compu-tational linguistics for extracting, classifying and interpreting millions of Short text expressions.

IV. EVALUATION MEASURES AND BENCHMARKS FOR SHORT TEXT ANALYSIS

Generally performance comparison is possible if experi-ments conducted are based on same dataset or corpus and test set. The same evaluation measures and parameter values should be used. The effectiveness of Short text Analysis (Similarity, Classification, Clustering) is usually measured in terms of the classic IR notions, like correlation coefficient, relative error, precision, recall, Fmeasure, accuracy, etc. [46].

Microsoft Paraphrase corpus is one of the common dataset used for Short text similarity measure. For experimental works in Short text analysis, eleven collections[1] (Micro4News, EasyAbstracts, SEPLN-CICLing, CICLing-2002-A, CICLing-2002-F, R4, R6, R8-, R8+, R8B and JCR) of Short texts with different levels of complexity have been provided. Recently Indian Forum for Information Retrieval Evaluation (FIRE-2011) have initiated a **SMS based FAQ Retrieval** task.

TABLE III
PERFORMANCE COMPARISON OF DIFFERENT SIMILARITY MEASUREMENT TECHNIQUES BASED ON MS PARAPHRASE CORPUS. *NM – NOT MENTIONED

Measure	Acc	Prec	Rec	F
Ramage et al.[4]				
Walk (Cosine)	0.687	NM	NM	0.787
Walk (Dice)	0.708	NM	NM	0.801
Walk (JS)	0.688	NM	NM	0.805
Abdulgader & Skabar[2]				
J&C	74.6	75.5	91.5	82.7
Path	73.2	73.9	92.4	82.1
Islam & Inkpen (2008) [5]				
STS	72.6	74.7	89.1	81.3
Michalcea et al.[3],Corpus based				
PMI-IR	69.9	70.2	95.2	81.0
LSA	68.4	69.7	95.2	80.5
Michalcea et al.[3],Wordnet based				
L&C	69.5	72.4	87.0	79.0
j&C	69.3	72.2	87.1	79.0
Baselines				
Vector-based	65.4	71.6	79.5	75.3
Random	51.3	68.3	50.0	57.8

The goal of this task is to find a question Q* from corpora of FAQs (Frequently asked questions) that best answers/matches the SMS query S. The data sets for the same is available.[2].

Table III list results of all experiments based on the Micosoft Paraphrase corpus. For the classification and clustering it is hard to find enough number of authors who used the same collection of dataset and training set in the same experimental conditions.

[1]http://sites.google.com/site/merrecalde/resources
[2]http://www.isical.ac.in/ clia/faq-retrieval/faq-retrieval.html

V. MAJOR ISSUES AND CHALLENGES

Although most of the prior works have tried to adress the common issues associated with Short texts and achieved fairly good results for selected domains, some other unresolved major issues are there.

Enriching data by using external sources are not feasible always and not suitable for real time applications. More than that enhancement of feature set may leads to the "curse of dimensionality". When feature set becomes very large, data becomes diffi-cult to visualize that results in higher model building time and also makes the classification or clustering slower. Traditional Bag-of-Words approach for representing fea-ture may not preserve the semantic context and it may give complete different connotation to the text message. Different methods are required to represent feature set.

Short text classification requires sufficient number of training examples to achieve high accuracy. It is often impractical and expensive to hand label a sufficient num-ber of training examples.

An accurate identification of semantic context requires better sentence structure analysis.

VI. CONCLUSION

The rich sources of Short text available in the web need to be analysed for various purposes like query analysis, classifying web search results, classifying or grouping blog posts and tweets, opinion mining etc. Short text messages are harder to analyse than long messages or documents. The major chal-lenge confronted by Shorttext analysis is the lack of content and context due to the data sparseness. This paper analyses different approaches to measure the similarity of Short texts and the methods used for classification and clustering. These approaches achieved fairly good results for the selected data set. This paper also summarizes the important issues associated with Short text analysis. In future more work is needed on further improving the techniques and to deal with the issues summarized in this paper.

REFERENCES

[1] Li YH, McLean D, Bandar ZA et al, "Sentence similarity based on semantic nets and corpus statistics.", IEEE Trans Knowl Data Engg 18, pp.1138-1150, 2006.

[2] Khaled Abdalgader, Andrew Skabar, "Short-Text Similarity Measurement Using Word Sense Disambiguation and Synonym Expansion.", LNAI 64649(Springer), pp.435-444, 2010.

[3] Mihalcea, R., Corley, C., Strapparava, C, "Corpus-based and Knowledge-based Measures of Text Semantic Similarity.", In 21st National Conference on Art. Int., vol. 1,2006, pp.775-780.

[4] Ramage, D., Rafferty, A., Manning, C, "Random Walks for Text Semantic Similarity.", In ACL-IJCNLP 2009, pp.2331.

[5] Islam, A., Inkpen, D, "Semantic Text Similarity using Corpus-based Word Similarity and String Similarity.", ACM Trans. on KDD 2(2), pp 1-25,2008.

[6] Okazaki N, Matsuo Y, Matsumura N et al, "Sentence extraction by spreading activation through sentence similarity.", IEICE Trans Inform Syst E86D(9), pp.1686-1694, 2003.

[7] Phan X, Nguyen L, Horiguchi S, "Learn to classify Short and sparse text and web with hidden topics from large-scale data collections.", In Proceedings of the 17th international conference on World Wide Web, 2008, pp.91-100.

[8] M. Sahami, T. Heilman, "A web-based kernel function for measuring the similarity of Short text snippets.", In Proc. of 15th International World Wide Web Conference,WWW 2006, May 2326, 2006, Edinburgh, Scotland, pp.377-386

[9] Xiaojun Quan, Gang Liu et al., "Short text similarity based on probabilistic topics.", Knowl Inf Syst, pp. 473-491, 2009.

[10] Metzler D, Dumais S, Meek C, "Similarity measures for Short segments of text.", In Proceedings of the 29th European conference on information retrieval (ECIR 2007), LNCS(Springer), Vol 4425,Berlin,2007, pp 16-27

[11] Yih W, Meek C, "Improving similarity measures for Short segments of text.", In Proceedings of twenty-second conference on artificial intelligence (AAAI-07), Vancouver, 2007, pp 1489-1494.

[12] Bollegala D, Matsuo Y, Ishizuka M, "Measuring semantic similarity between words using Web search engines.", In Proceedings of the 16th international conference on World Wide Web, 2007, ACM Press, New York, pp. 757-766.

[13] S. Zelikovitz and H. Hirsh, "Improving Short-Text Classification using Unlabeled Data for Classification Problems.", In Proceedings of the Sev-enteenth International Conference on Machine Learning, 2000, pp.1191-1198.

[14] Qiang Pu and Guo-Wei Yang, "Short-Text Classification Based on ICA and LSA.", LNCS(Springer) 3972, pp 265-270, 2006.

[15] Andreas Heb, Philipp Dopichaj and Christian Maab, "Multi-Value Classification of Very Short Texts.", In Proceedings of the 31st annual German conference on Advances in Artificial Intelligence, Springer-Verlag Berlin, 2008, pp.70-77.

[16] Bharath Sriram, David Fuhry, Murat Demirbas, "Short Text Classification in Twitter to Improve Information Filtering.", ACM SIGIR, Geneva, Switzerland, 2010, pp.841-842.

[17] Evan Wei Xiang, Qiang Yang, "Knowledge Base assisted Text Categorization.", ACM-HK Student Research and Career Day, 2009.

[18] Sarah Zelikovitz, Haym Hirsh, "Using LSI for text classification in the presence of background text.", In Proceedings of the tenth international conference on Information and knowledge management(CIKM), 2001, pp. 113-118.

[19] S. Zelikovitz, Transductive LSI for Short Text Classification Problems, In Proceedings of the Seventeenth International FLAIRS, 2004.

[20] HUANG Wei, Li Shan-fei, Tan Yue-jin, Gao Bing,"Association Rules Based Short Text Feature Extension.", IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.10, pp. 227-230, Oct. 2009.

[21] Zitao Liu, Wenchao Yu, Wei Chen, Shuran Wang, Fengyi Wu, "Short Text Feature Selection for Micro-blog Mining.", In proceedings of Com-putational Intelligence and Software Engineering (CiSE), Wuhan,2010, pp. 1-4.

[22] Zhou Faguo, Zhang Fan, Yang Bingru, Yu Xingang, "Research on Short Text Classification Algorithm Based on Statistics and Rules.", In proceedings of third International Symposium on Electronic Commerce and Security(2010), pp. 3-7.

[23] Banerjee S, Ramanathan K, Gupta A, "Clustering Short text using Wikipedia.", In Proceedings of the 30th international ACM SIGIR conference on research and development in information retrieval (SIGIR) ,2007, pp. 787-788.

[24] Xia Hu, Nan Sun, Chao Zhang, Tat-Seng Chua , "Exploiting internal and external semantics for the clustering of Short texts using world knowledge.", In Proc. CIKM Hong Kong, China, Nov. 2009, pp. 919-928.

[25] Zamir O, Etzioni O, "Grouper: A dynamic clustering interface to web search results.", In Proceedings of the 8th international conference on World Wide Web, 1999 , pp.1361-1374.

[26] Zamir O, Etzioni O,"Web document clustering: a feasibility demonstration.", In Proceedings of the 21th international ACM SIGIR conference on research and development in information retrieval (SIGIR) ,1998, pp. 46-54.

[27] Hui He, Bo Chen, Weiran Xu, Jun Guo, "Short Text Feature Extrac-tion and Clustering for Web Topic Mining.", In Proceedings of Third International Conference on Semantics, Knowledge and Grid, 2007, pp. 382-385.

[28] Kummamuru K, Lotlikar R, Roy S, Singal K, Krishnapuram R, "A hierarchical monothetic document clustering algorithm for summarization and browsing search results.", In Proceedings of the 13th international conference on World Wide Web, 2004, pp. 658-665.

[29]. Makagonov, P., Alexandrov, M., Gelbukh, A, "Clustering abstracts instead of full texts.", In: Proceedings of TSD-2004, LNAI, vol. 3206, pp129-135.

[30] Alexandrov, M., Gelbukh, A., Rosso, P, "An approach to clustering abstracts.", In: Proceedings of NLDB-05, LNCS(Springer), vol. 3513, pp. 8-13, 2005.

[31] Pinto, D. "On Clustering and Evaluation of Narrow Domain Short-Text Corpora.", PhD Dissertation. Universidad Politencnica de Valencia, 2008

[32] Pinto D., Rosso P., Jimnez H, "A Self-enriching methodology for clustering narrow domain Short texts.",The Computer Journal, 54(7), pp. 1148-1165, 2011.
qq.

[33] Li Yang, Adnan Rahi, "Dynamic Clustering of Web Search Results.", LNCS 2667(Springer), 2003, pp. 153-15.

[34] Hua-Jun Zeng, Qi-Cai He, Zheng Chen, Wei-Ying Ma, Jinwen Ma, "Learning to Cluster Web Search Results.", ACM SIGIR04,South Yorkshire, UK, 2004, pp.210-217.

[35] Wang X, Zhai C, "Learn from web search logs to organize search re-sults.", In Proceedings of the 15th international ACM SIGIR conference on research and development in information retrieval,2007, pp. 8794

[36] Xingliang Ni, Xiaojun Quan, Zhi Lu, Liu Wenyin, Bei Hua, "Short text clustering by finding core terms.", Knowledge and Information Systems,Springer,27, pp.345-365, 2011.

[37] Hsinchun Chen, David Zimbra, "AI and Opinion mining.",IEEE Intelligent Systems,vol. 25 no.4, 2010.

[38] Huifeng Tang, Songbo Tan, Xueqi Cheng, "A survey on sentiment detection of reviews.", Expert Systems with Applications,ELsevier, 36, pp. 10760-10773, 2009

[39] Bing Liu, "Sentiment Analysis and Subjectivity.", Handbook of Natural Language Processing, Second Edition, 2010

[40] Yelena Mejova, "Sentiment Analysis: An Overview, Comprehensive Exam Paper.", 2009.

[41] Yee W. LO, Vidyasagar, "A Review of Opinion Mining and Sentiment Classification Framework in Social Networks.", 3rd IEEE International Conference on Digital Ecosystems and Technologies, 2009, pp 396-401.

[42] Thelwall, M., Buckley, K et. al. "Sentiment strength detection in Short

[43] informal text.", Journal of the American Society for Information Science and Technology, 61(12), pp.25442558, 2010.

[44] L. Xu, A. Krzyzak, E. Oja, "Rival Penalized Competitive Learning for Clustering Analysis, RBF Net and Curve Detection.", IEEE Transactions on Neural Networks, pp.636-649, 1993

[45] David M. Blei, Andrew Y. Ng, Michael I. Jordan, "Latent Dirichlet Allocation.", Journal of Machine Learning Research 3, pp.993-1022, 2003

[46] Dolan, W., Chris Quirk, C., Brockett, C.V, "Unsupervised Construc-tion of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources.", In proceedings 20th International Conf. on Computational Linguistics,2004, pp. 350-356.

[47] J. Han, M. Kamber, "Data Mining: Concepts and Techniques.", Morgan Kaufmann Publishers(Elsevier), San Francisco, 2001